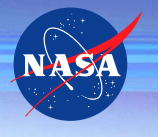

National Aeronautics and Space Administration
NASA
Exploring the Overlap of High Performance Computing and Big Data
Daniel Duffy
High Performance Computing Lead
NASA Center for Climate Simulation
Goddard Space Flight Center
www.nasa.gov

Agenda



Introduction to the NCCS

- Dr. Phil Webster – Data-Centric Science at the NASA Center for Climate Simulation

Comparison of HPC and Internet Technologies

Science Cloud

- Hoot Thompson Presentations in the RedHat and Mellanox Booths

Data and Data Services

- Laura Carrier - Advancing Research and Applications with NASA Climate Model Data

MERRA A/S

- Glenn Tamkin – MERRA Analytic Services: Orchestrating Big Data with Climate Analytics-as-a-Service

Accelerator Work

- Craig Pelissier – Climate Dynamics on the Intel Xeon Phi (Craig Pelissier)
- Doris Pan – Porting Earth Science Applications to Next-Generation Intel Xeon Phi Coprocessors

People That Make NCCS Successful

NASA Center for Climate Simulation (NCCS)

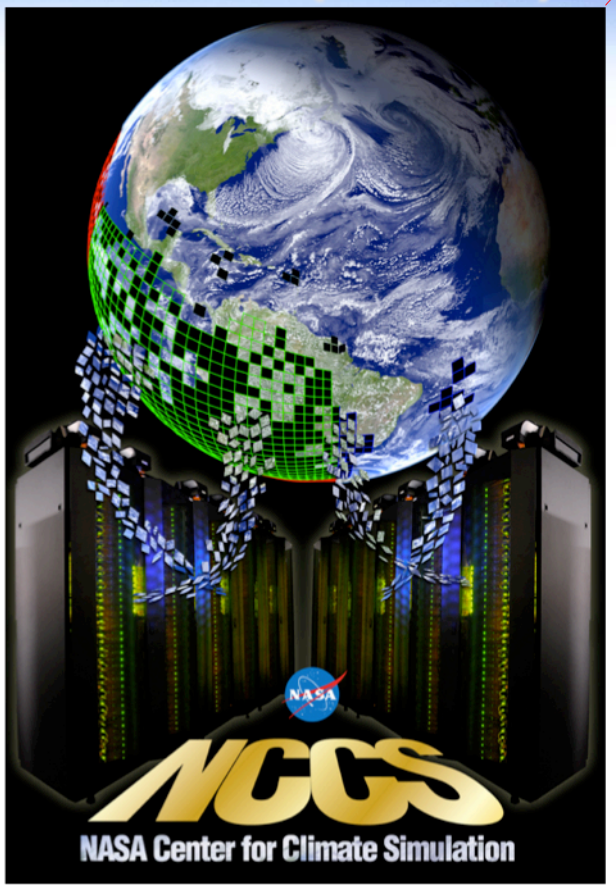
Funded by the Science Mission Directorate

- Located at the Goddard Space Flight Center (GSFC)

Provides an integrated high-end computing environment designed to support the specialized requirements of Climate and Weather modeling.

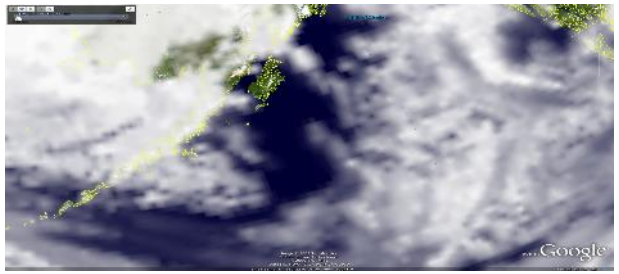
- State-of-the-art high-performance computing, data storage, and networking technologies
- Advanced analysis and visualization environments
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services

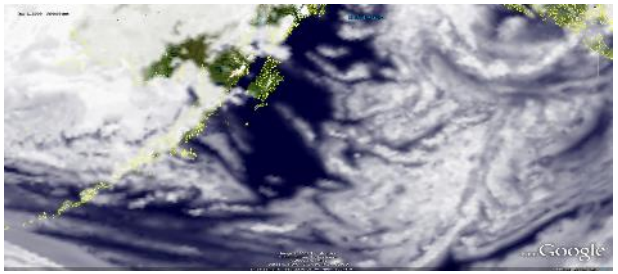
http://www.nccs.nasa.gov

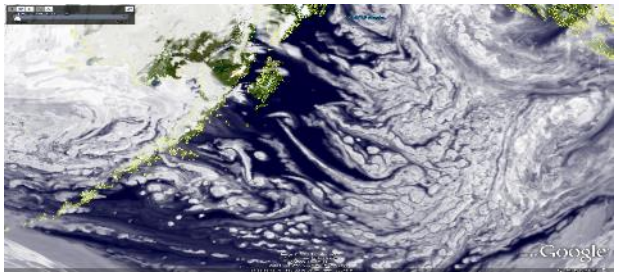


Increasing Global Model Resolution

Current Operations | **Cloud-Permitting** | **Cloud-Resolving**







Requirement	Current Operations	Cloud-Permitting	Cloud-Resolving
Number of Cores	100's	300,000	10,000,000
Resolution	27 KM	10 KM to 3 KM	1 KM or Finer
Number of Racks	1 Rack	234 Racks	7,800 Racks
Total Power	20 KW	4. 7 MW	100 MW

Assuming current compute technology (Intel SandyBridge), the computer needed to run a cloud-resolving model does not exist today and would require entirely too much power. A different approach is needed – adoption of low-power highly parallel processors.

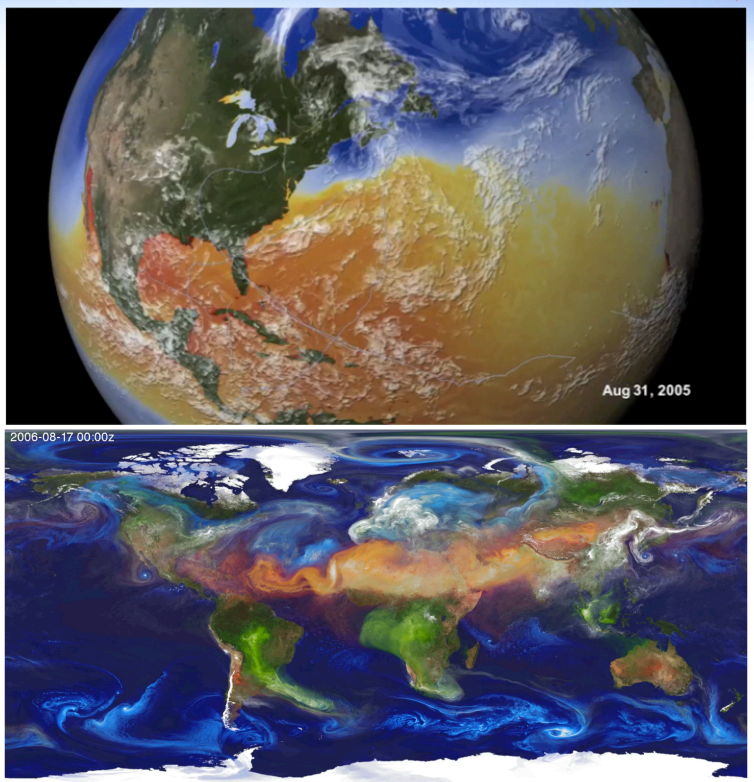
GEOS-5 Global Mesoscale Simulation

Climate Science is Data Intensive

- As systems capabilities increase, the amount of data generated by the science is growing at an alarming rate
- Providing a challenge for file systems and data analysis

GEOS-5 Nature Run

- 2-year Nature Run at 7.5 KM resolution
- 3-month Nature Run at 3.5 KM resolution
- Will generate about 4 PB of data (compressed)
- To be used for Observing System Simulation Experiments (OSSE's)
- All data to be publically accessible



Top: 10-km GEOS-5 meso-scale simulation for Observing System Simulation Experiments(OSSEs)
Bottom: The Goddard Chemistry Aerosol Radiation and Transport (GOCART) model,
Courtesy of Dr. Bill Putman, Global Modeling and Assimilation Office (GMAO), NASA Goddard Space Flight Center.

NCCS Computational Growth

Continue to deploy scalable units into the Discover Cluster

Truly a hybrid system

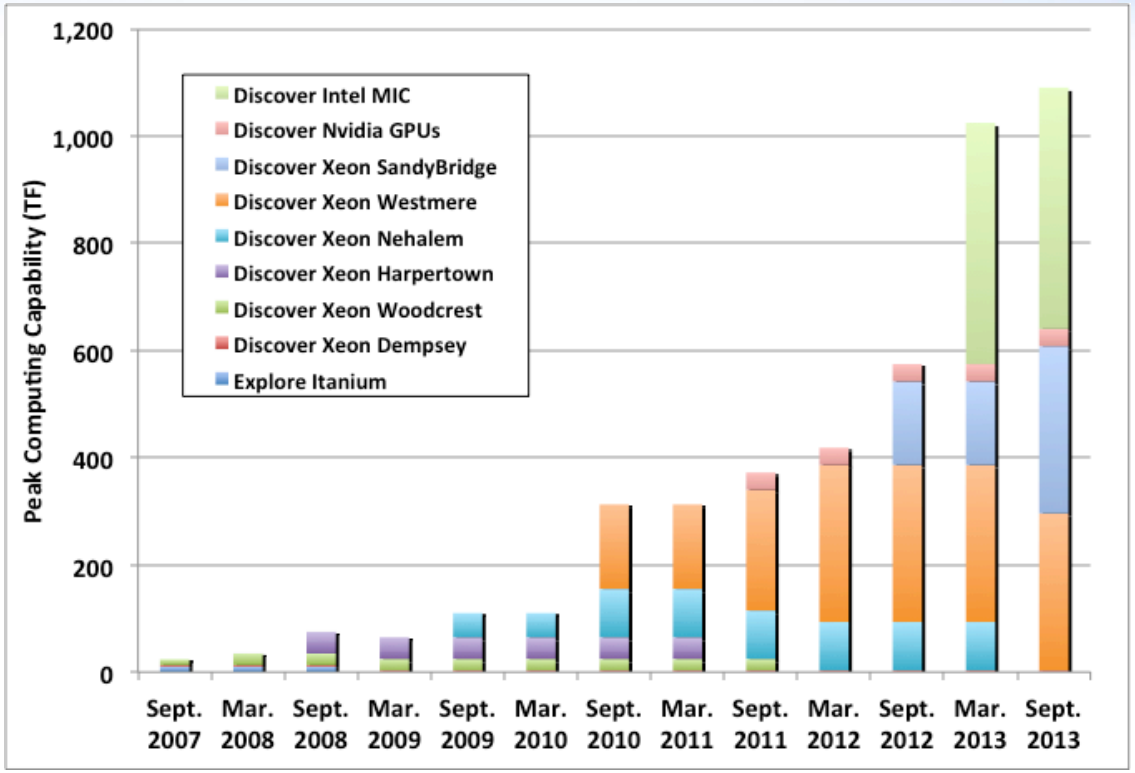
- Xeon only nodes
- Nodes with GPUs
- Nodes with Intel Phi

Major milestone for the NCCS in 2012

- Exceeded 1 PF Peak!

Growth over the last 10 years

- 300x increase in compute
- 2,000x increase in storage



2013 Compute Upgrade – SCU9

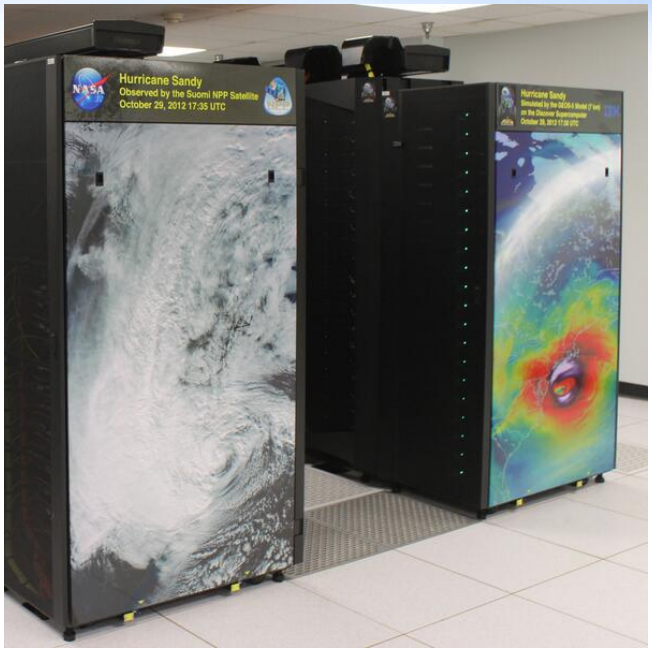
Scalable Unit Number 9

- Installed in Summer of 2013
- IBM iDataPlex
- 480 compute nodes
- Intel Xeon SandyBridge Processors
- 64 GB of RAM
- FDR Infiniband

Computational Capability

- Peak: 159,744 Gflops

Capable of additional Intel Phi coprocessors or NVIDIA GPUs



Come see Bruce Pfaff's talk *Technology Updates at the NASA Center for Climate Simulation* **at the NASA booth for more information!**

Data Centric HPC, Big Data and IT Environment

2006

Data Sharing and Publication

- Capability to share data & results
- Supports community-based development
- Data distribution and publishing

Code Development

- Code repository for collaboration
- Environment for code development and test
- Code porting and optimization support
- Web based tools

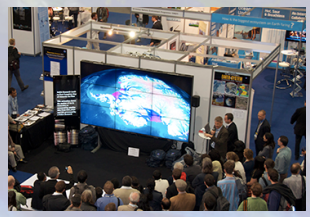
User Services

- Help Desk
- Account/Allocation support
- Computational science support
- User teleconferences
- Training & tutorials

DATA Storage & Management

Analysis & Visualization

- Interactive analysis environment
- Software tools for image display
- Easy access to data archive
- Specialized visualization support

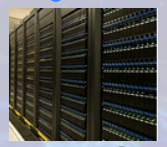


Data Transfer

- Internal high speed interconnects for HPC components
- High-bandwidth to data center users
- Multi-gigabit network supports on-demand data transfers

Global file system enables data access for full range of modeling and analysis activities

Security



HPC Computing

- Large scale HPC computing
- Comprehensive toolsets for job scheduling and system monitoring

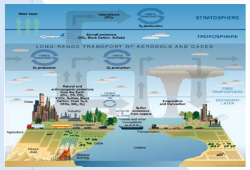
Data Archival and Stewardship

- Large capacity storage
- Tools to manage and protect data
- Data migration support

Evolution to a Data Services Centric Environment

Data

HPC Models
- GEOS 5
- ModelE
- WRF



Observations
- Ground Based
- Satellite
- In Situ

Reanalysis
- MERRA
- NOAA
- Others

HPC Computing and Storage
- NASA NCCS
- NOAA
- Others

Analytics

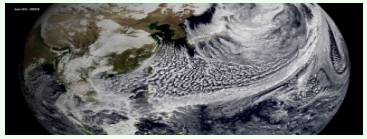
Data Services

Moving beyond just a file system and a storage repository.

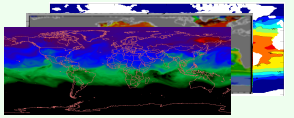
NCCS and Data Services Projects
- Dali Analysis Nodes
- vCDS
- Hadoop (HDFS)
- Merra Analytic Service
- Earth System Grid
- Web Portals

Discovery

Modelers/ Scientists



Downstream Users
- Agriculture
- Water Management
- Health
- Famine Prediction



Commercial
- Insurance/Reinsurance
- Commodity Trading

Public/Citizen Scientists

Data Management System
iRODS based management of federated data sets

Data Analysis and Analytics Technology Gap

Archive

Archive
~1 PB of Disk
~35 PB of Tape
Optimized for long term storage, typically slower storage designed for streaming reads and writes

Leads to Un-optimized Practices:
Users perform data analysis straight from the archive and complain that it is too slow.

Very Large Performance Gap

Specifically for Data Analysis, Analytics, and Visualization of large scale data

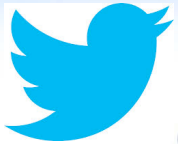
What technologies can we use to help bridge this gap?

Large Scale Compute

Discover Cluster
>1 PF Peak
~18 PB of Disk
Optimized for large scale simulations with fast storage designed for streaming applications

Leads to Un-optimized Practices:
Users analyze large data sets through a series of many small blocks reads and writes and complain that it is too slow.

Shifting Technologies Toward Data



High Performance Computing

- Shared everything environment
- Very fast networks; tightly coupled systems
- Cannot lose data
- Big data (100 PBs)
- Bring the data to the application
- Large scale applications (up to 100K cores)
- Applications cannot survive HW/SW failures
- Commodity and non-commodity components; high availability is costly; premium cost for storage

Object Storage
MapReduce
Hadoop
Cloud
Open Stack
Virtualization
Accelerators

Large Scale Internet

- Examples: Google, Yahoo, Amazon, Facebook, Twitter
- Shared nothing environment
- Slow networks
- Data is itinerant and constantly changing
- Huge data (Exabytes)
- Bring the application to the data
- Very large scale applications (beyond 100Ks)
- Applications assume HW/SW failures
- Commodity components; low cost storage

HPC Science Cloud

Adjunct to Discover hosted science processing

- Lower barrier to entry for scientists
- Customized run-time environments
- Code validation against older system images

Expanded customer base

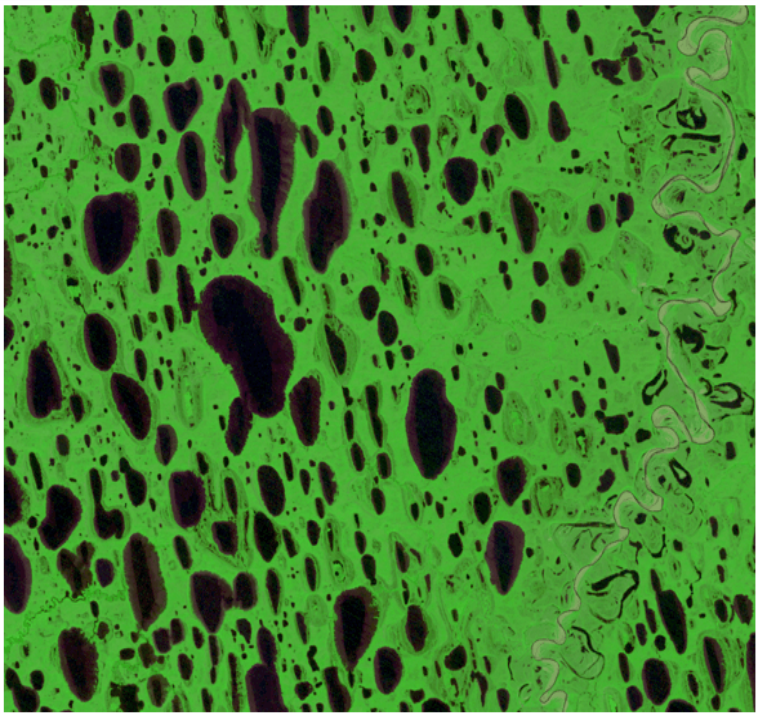
- Science Data Processing (e.g. ABoVE)
- Temporal processing campaigns (e.g. ifloods)

What is different than a commodity could?

- Come close to matching HPC levels of performance
- Node-to-node communication critical – high speed, low latency, RDMA
- Shared, high performance file system mandatory
- Management and rapid provisioning of resources – cluster formation

Biggest obstacle – performance loss in virtualized space

Representative Science Cloud Application



0 km 5

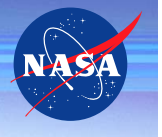
Representative Landsat image, false color composite, from near Barrow, AK

Decadal Water Products for ABoVE

- Mapping the water in the High Northern latitudes for the past three decades prior to the Arctic Boreal Vulnerability Experiment (ABoVE) field campaign
- Time series of Landsat images at 30 meter spatial resolution at 3 epochs (1991, 2001, and 2011)
- Calculate the maximum, minimum, and average condition of each lake or pond 1 ha or larger for each epoch

Working with Dr. Mark Carroll (Goddard Space Flight Center

Optimizing Communications for Virtual Systems

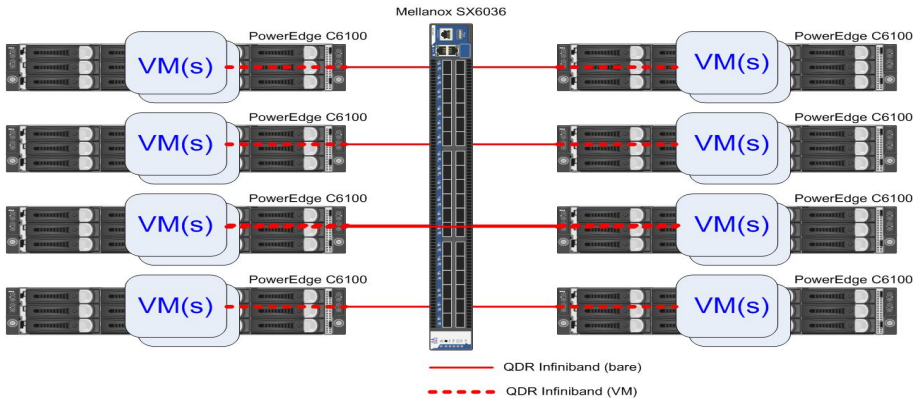


Virtualized Inifinband

- Using Single Root I/O Virtualization (SR-IOV)
- VM tuning – hugepages and NUMA aware scheduling of processes

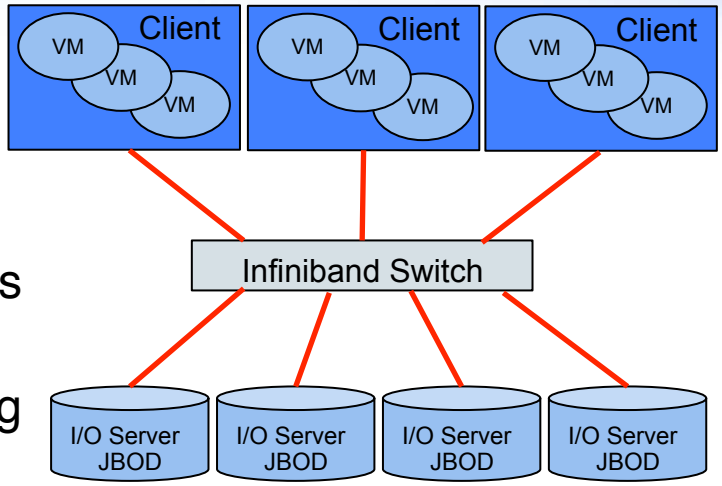
Current Test Bed

- Eight node Dell C6100 proof of concept cluster
- Each node has 12 cores and 24 GB of RAM
- QDR Infiniband interconnect
- Gluster File System



Working closely with both RedHat and Melllanox on this work.

File System for Science Cloud
GlusterFS Operational Prototype



Recently acquired 960TB raw storage

- Four I/O servers
  - IB frontend
  - SAS backend
- Four 60-bay JBODS populated with 4TB drives
  - One per I/O server
- Carving out various Gluster volumes for testing

Bare metal Gluster clients

- Connected to I/O servers using IB/RDMA

VM Gluster clients

- Connected to I/O servers using virtualized IB/RDMA or IPoIB
- Currently only using one I/O server (just got the storage two weeks ago!)

Progress on ABoVE Project

Approximately 25,000 Landsat images to process

- Python script takes about 7 minutes to process a single image
- Would take approximately 4 months on the scientist's desktop

Current Progress – Ongoing as we speak

- 8 VM's being used concurrently
- Gluster file system served out IPoIB
- Over 3,500 scenes processed in just a few days
- Expect the entire 25,000 images to take about a week by the time we are done

Future Work on the HPC Science Cloud

Scaling Out

- Low level benchmarks show good results
  - Streams, Linpack, NPB
- Want to run additional applications, including
  - GEOS-5 Dynamical Core, WRF

File System

- Issues with Gluster RDMA
- Need to get those resolved for best performance
- Scale out the storage across multiple Gluster targets

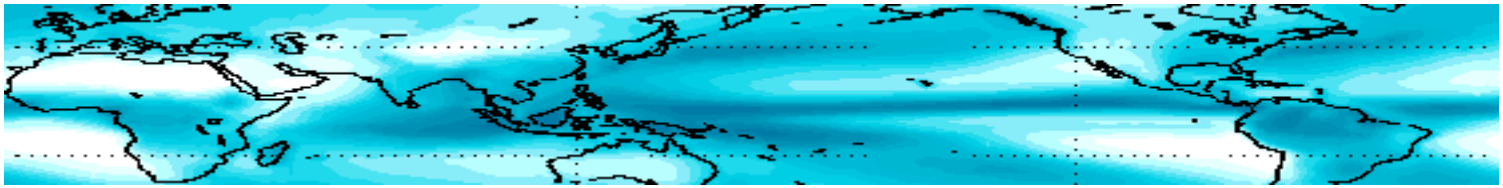
For More Information

- Hoot Thompson will be speaking at both the Mellanox and RedHat booths this week about this topic

Modern Era Retrospective-Analysis for Research and Applications (MERRA)

Format	NetCDF
Product Resolutions	Native ($^1/_2{}^o$ x $^2/_3{}^o$, using model conventions) Reduced ($1^1/_4{}^o$ x $1^1/_4{}^o$, dateline-edge, pole-edge) Reduced FV (1^o x $1^1/_4{}^o$, using model conventions)
Reference	http://gmao.gsfc.nasa.gov/merra/
Download	http://disc.sci.gsfc.nasa.gov/misc/data-holdings

Retrospective-analyses (or reanalyses) have been a critical tool in studying weather and climate variability for the last 15 years. Reanalyses blend the continuity and breadth of the output data of a numerical model with the constraints of vast quantities of observation data. The result is a long-term continuous data record. MERRA was developed to support NASA's Earth science objectives, by applying the state-of-the-art GMAO data assimilation system that includes many modern observing systems (such as EOS) in a climate framework.



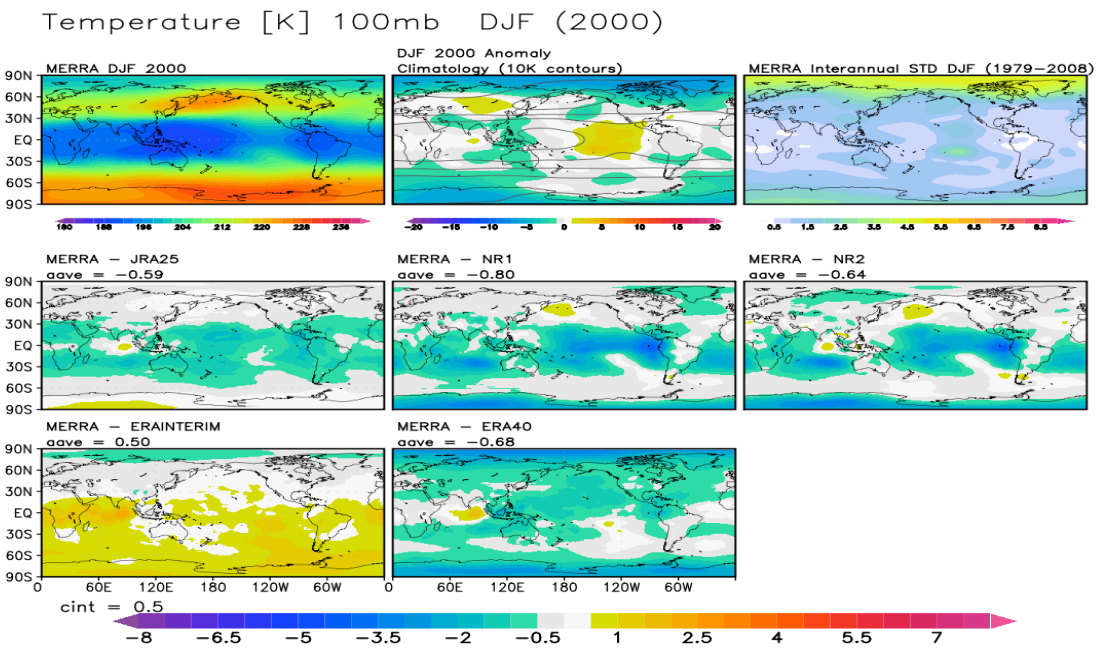
Simplified Operation on MERRA Data

Create a time based average over the monthly means for specific variables

This example shows a seasonal average of temperature for the winter of 2000

Canonical Operations

- Max, Min, Stdev, Count, Sum, Var
- Open specific files
- Read NetCDF header information
- Extract parameter
- Close file
- Repeat for all relevant files
- Average the parameter



This type of operation should translate well to a map reduce function.

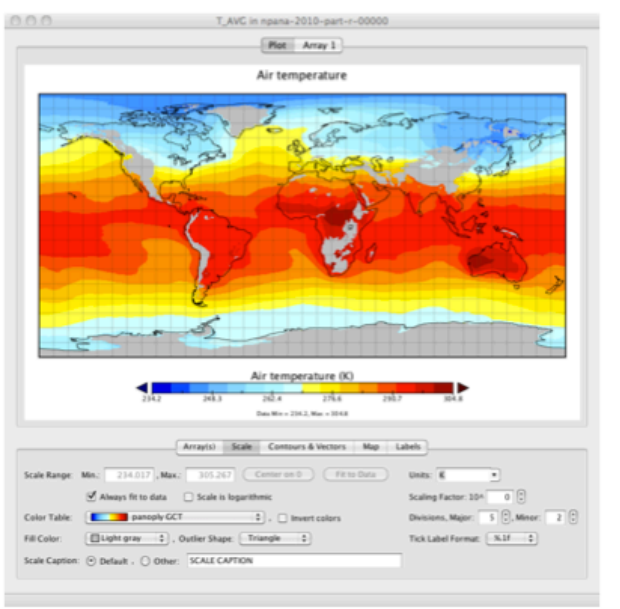
MERRA Analytic Service

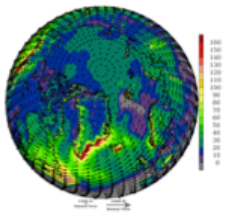
Cyber Infrastructure Resource

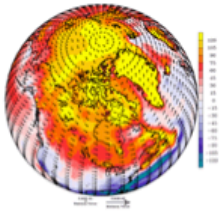
- Combines iRODS data grid and Hadoop MapReduce capabilities to server MERRA analytics products
- Stores MERRA reanalysis data in an HDFS to enable parallel, high-performance, storage-side data reductions
- Provides a library of commonly used spatiotemporal operations (canonical operations) that can be stitched together to enable higher order analyses

Funded under the NASA 2013 A.40 ROSES Solicitation

- Dr. John Schnase (PI) and Dr. Daniel Duffy (Co-PI)
- Special thanks to Dr. Tsengdar Lee (NASA HQ) and all participants in the A.40 work.







See Glenn Tamkin at the NASA Booth to learn more about the *MERRA Analytic Services: Orchestrating Big Data with Climate Analytics-as-a-Service*

MERRA A/S Cluster

- 36 node Dell cluster, 576 Intel 2.6 GHz SandyBridge cores, 1300 TB raw storage, 1250 GB RAM, 11.7 TF theoretical peak compute capacity.
- FDR Infiniband network with peak TCP/IP speeds >20 Gbps.

Component	Configuration
Node	Dell R720
Processor Type	Intel SandyBridge
Processor Number	E5-2670
Processor Speed	2.6 GHz
Cores per Socket	8
Number of Sockets	2
Cores per Node	16
Main Memory	32 GB
Storage	12 by 3 TB drives = 36 TB RAW
Interconnect	Mellanox MT27500 FDR IB
Operating System	CentOS 6.3
Kernel	2.6.32-279.5.1
Hadoop	0.20.2
Java	1.6.0_24

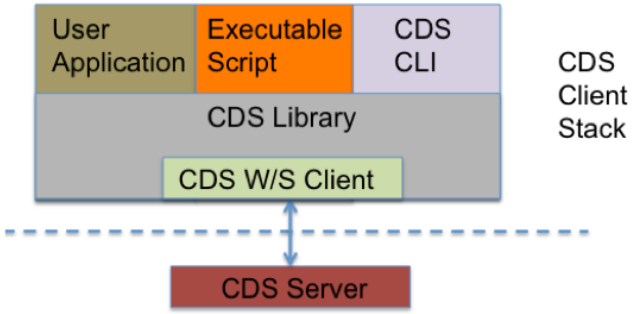
NASA Climate Data Services Application Programming Interface

Starting point for the development of a NASA Climate Data Services Application Programming Interface (CDS-API)

- CDS client stack distributed as a software package
- Could be used to be build a cloud service (SaaS) or cloud image

Climate Analytics-as-a-Service (CAaaS)

- This approach to an API design focuses on specific analytic requirements for climate sciences



Options for using the MERRA A/S

- Reference Model
- Library
- Web Service Client
- Command Line Interface
- Applications
- Scripts

NCCS Phi Installation One Year Ago

Scalable Unit Number 8

- Installed in October 2012
- 480 compute nodes with Intel Phi co-processors

Computational Capability

- Peak: 620,544 Gflops
- 221 KW Total Power

Target Applications

- Goddard Earth Observing System (GEOS-5)
- WRF
- Earth's Magnetic Field – Hall2D MHD
- Ice Melting Code
- Gravity Field (GRAIL Mission)

For More Information:

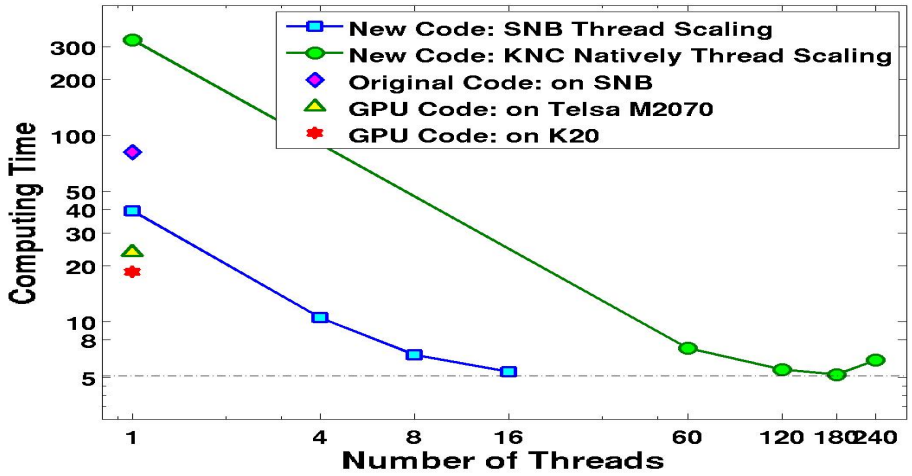
Craig Pelissier (NASA Booth Demonstration) – *Climate Dynamics on the Intel Xeon Phi*

Doris Pan (NASA Booth Presentation) – *Porting Earth Science Applications to Next-Generation Intel Xeon Phi Coprocessors*

Promising Results – Hall2D MHD

Investigate the properties of magnetic reconnection at Earth's dayside magnetosphere

Goal: Port the entire code to MIC using native/symmetric model



- Used OpenMP to exploit a large amount of concurrency in the application
- Single node comparisons for Xeon, Phi, and GPU
- Optimizations in the application benefitted both the Phi and the Xeon
- Application scales well on the Phi up through 180 threads and matches the Xeon node performance

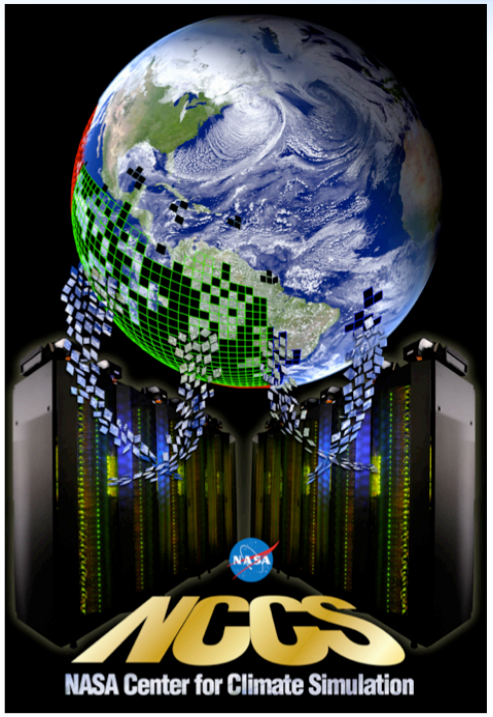
Lessons Learned So Far

Low Barrier to Entry
- Can get applications running on the Phi relatively quickly
- Intel tools work well (compilers, debuggers, and performance tools)

Phi Optimizations
- Benefit the overall performance of the applications on the Phi AND on the Xeon
- Optimization is a constantly moving target

Lots of Work To Do
- Get full applications running in symmetric mode
- Optimize the system

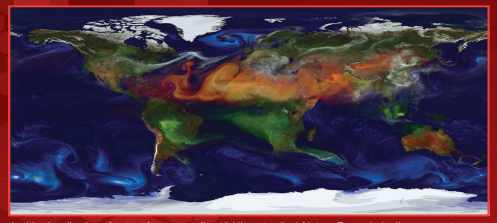
Concurrency is the Future
- These types of optimizations will benefit our applications on future platforms
- Make the investment now



Thank You

National Aeronautics and Space Administration

Exploring the Overlap of High-Performance Computing and Big Data for Climate



In this visualization of output from an earlier 10-kilometer (km) Nature Run, winds disperse vast quantities of aerosols around the world—dust (red), sea salt (blue), sulphate (white), and black and organic carbon (green). A new Nature Run will increase global resolution to 7.5 km for 2 years and 3.5 km for 3 months. *William Putman, NASA/Goddard*

A 7-km global simulation not only produced an accurate track for Hurricane Sandy but also captured fine-scale details of the storm's changing intensity and winds. In this visualization, near-surface wind speeds range from dark blue (10 miles per hour) to light purple (100 miles per hour). *William Putman, NASA/Goddard*

The NASA Center for Climate Simulation (NCCS) is constantly evaluating new and innovative technologies for potential use within its high-performance computing (HPC) environment. Traditionally, these technologies have focused on how to make models execute faster and scale to larger core counts. As NCCS has integrated new HPC capabilities, the data generated by the super-high-resolution climate models has grown exponentially.

This confluence of HPC and data requirements has created a new context for evaluating technology. NCCS has developed a strategic view of information technology based on key architectural aspects of a mix between HPC and large Internet service providers.

Out of this new context, NCCS has explored a number of different technologies to enhance its capabilities to provide both HPC and high-end data services for science. These include:

- Creation of a prototype science cloud for use in satellite data processing and engineering.
- Remote visualization capabilities for large-scale climate data.
- Data services, including the creation of a Climate Model Data Services Application Programming Interface (CDS API).

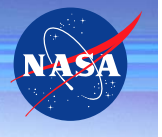
Evaluation and testing of these technologies occurs within the NCCS architectural context. Current challenges include gaps between HPC and big data, which will be addressed by upcoming applied research.

Daniel Duffy, NASA Goddard Space Flight Center

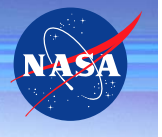
SUPERCOMPUTING
SCIENCE MISSION DIRECTORATE

www.nasa.gov

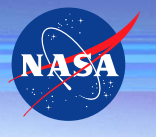
Introduction to the NCCS



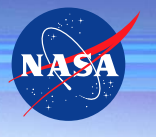
Technology Gaps



Science Cloud



MERRA A/S



Accelerator Work